April 2021

NHLBI TOPMed Program

Ethical, Legal, and Social Issues (ELSI) Committee Summary: Stratified frequencies

Contents

Executive Summary 2
- Summary of recommendations for TOPMed 2

Background 3

Issue statement 3

Points to consider 4
- Benefits 4
  - Researchers to identify populations for replication or “look up” studies 4
  - Researchers to compare allele frequencies for validation of findings or overall QC 4
  - Clinical researchers and clinicians to assess variant pathogenicity 5
- Concerns 5
  - Re-identification 6
  - Reputational harm 6
  - Reification of race 6
  - Consent issues 7
- Considerations for implementation 7
  - Figure 1: Visualization of approaches for computing stratified allele frequencies. 8

Conclusion and Recommendations 10

References 13

Appendices 15
- Advantages and disadvantages to grouping approaches 15
  - Self-reported racial and/or ethnic categories 15
  - Predominant ancestry groups 15
- TOPMed effective sample size 16
  - Figure S1. Effective sample size 16
  - Figure S2. Ancestry proportions 16
- Precedent from prior and existing genomics resources 17
  - 1000 Genomes Project 17
  - gnomAD 17
  - ALFA (Allele Frequency Aggregator) 18
  - TOP-LD 19

Document history 19

Executive Summary

This summary document offers points to consider for TOPMed study investigators as they, or their advisory bodies, decide whether their study will contribute to stratified frequencies in the TOPMed BRAVO server. To date, only TOPMed-wide allele frequencies have been presented in BRAVO, i.e., frequencies are not stratified by genetic ancestry, race/ethnicity, study membership, or any other demographic or genetic features of TOPMed participants. However, stratified frequencies are commonly requested by BRAVO users and may further increase the scientific benefit of the resource.

There are several potential benefits and concerns in adding stratified frequencies. Benefits include allowing researchers to identify populations for replication or “look up” studies, enabling researchers to compare population-specific frequencies to validate findings, and facilitating assessment of variant pathogenicity in clinical contexts. Concerns include potential reidentification (determining whether or not an individual participated in a given research study based on allele frequencies and individual-level genetic data), reputational harm of associating stigmatizing variants with specific groups, potential reification of race as a biological rather than a socio-political category, and whether presenting stratified frequencies aligns with participants’ consents.

To mitigate these concerns, we recommend a novel approach of estimating ancestry-specific allele frequencies using a statistical deconvolution method and based on local genetic ancestry inference. Notably, this method does not require grouping individuals by either predominant global ancestry or race/ethnicity and therefore mitigates reidentification and other concerns because the mixture distribution of ancestral allele frequencies varies across the genome.

Summary of recommendations for TOPMed

1. Provide ancestry-specific allele frequencies by statistical deconvolution, in addition to TOPMed-wide frequencies.
2. Individual studies and cohorts may choose to additionally participate as a unique strata if desired.
3. Study PIs and representatives should carefully consider participant consent when deciding whether to contribute to stratified allele frequencies in BRAVO.
4. TOPMed stratified frequencies should be transparently and consistently described and used across resources and publications.
5. The TOPMed program should revisit and revise the definition and composition of strata in TOPMed as needed moving forward.

Background

TOPMed variant information (e.g., variant names, allele frequencies, and annotations) is currently publicly available in resources such as the TOPMed Informatics Research Center's BRAVO server, NCBI dbSNP, and, more recently, FAVOR. Deposited data are from TOPMed studies where study investigators have explicitly agreed to contribute to these public-facing resources. To date, only TOPMed-wide allele frequencies have been presented, i.e., frequencies are not stratified by genetic ancestry, race/ethnicity, study membership, or any other demographic or genetic features. However, stratified frequencies are commonly requested by BRAVO users and will potentially further increase the scientific benefit of the resource. The numbers of TOPMed studies and study participants have also increased considerably since BRAVO was first developed, meriting a revisiting of concerns about stratification previously raised by the initial cohort of TOPMed studies.

The purpose of this summary document is to offer points for TOPMed study investigators to consider as they, or their advisory bodies, decide whether their study will contribute to stratified frequencies in BRAVO. This document also proposes a novel approach of defining ancestry-specific frequencies based on local ancestry inference, which alleviates many of the concerns with prior approaches of categorizing individuals into fixed groups based on either genetic or demographic information.

Uses of allele frequencies

Allele frequencies are used by scientists and laboratories in the following ways:

- Validation of a research result or finding against a known parameter. For example, when a trait-variant association is detected, a researcher may want to verify that the information about the variant is correct, e.g., that the estimated allele frequency matches known frequencies.
- Identification of a population in which a variant is relatively common in order to design a replication study.
- Variant interpretation in a clinical context, i.e., as evidence for classifying variants as "pathogenic," "benign," or "uncertain significance." Frequency can be obtained from an aggregate dataset regardless of subpopulations, or within subpopulations.
- Population genetics research, where allele frequencies can be used to study population history and evolutionary processes.

In each of these examples, the utility of allele frequencies may increase when based on defined population subsets or strata, as explored below.

Issue statement

Presenting stratified allele frequencies on BRAVO would likely increase the scientific benefit of the resource yet may introduce some concerns for contributing studies. Additionally, there are

multiple approaches to creating or defining strata, each of which has advantages and disadvantages.

Points to consider

Benefits

TOPMed is one of the largest collections of whole genome sequences to date, with 78.7% of variants discovered not previously reported in dbSNP (Taliun et al, 2021). Indeed, TOPMed data in the BRAVO server may be the only public resource with a record of a given variant of interest. Making TOPMed variant information available to the scientific community via resources such as BRAVO facilitates broad use. While TOPMed-wide frequencies provide valuable information, further stratifying frequencies would provide additional benefits for genetic epidemiology, bioinformatics, and clinical investigation, by enabling:

1. Researchers to identify populations for replication or “look up” studies

Genome-wide association studies (GWAS), based on either common or rare/infrequent variants, result in a list of variants that are potentially associated with an outcome of interest (“putative associations”). Due to a large number of potential false positive associations, a variant is accepted as likely associated with an outcome only once a stronger evidence of association (compared to a single-study discovery) is established. A standard practice is to replicate the variant association in an independent population. Replication testing needs to be performed in a study population in which the variant exists. Thus, researchers may consult TOPMed variant frequencies to identify opportunities for replication testing. Ideally, a researcher would be able to either identify a specific study to contact for replication testing, or identify a broader population category, such as a genetic ancestry or race/ethnic group, that will ultimately enable the identification of a specific study or biobank-based population for replication. However, it is difficult to develop a specific replication plan when presented with only a TOPMed-wide frequency that comprises dozens of studies and heterogeneous ancestries. Presenting more granular frequencies would guide the researcher towards the population(s) needed to pursue replication and validation.

2. Researchers to compare allele frequencies for validation of findings or overall QC

In addition to identifying populations in which to pursue follow-up validation studies, stratified frequencies can aid other aspects of GWAS. For example, when a trait-variant association is detected, researchers may want to verify that the information about the variant is correct, e.g., that the estimated allele frequency matches known frequencies. Comparing allele frequencies between one’s study population and publicly available reference populations can also aid general quality control (QC) of genotyping array or imputation data, e.g., substantially differing frequencies can indicate issues with genotyping technologies or annotation (e.g., strand orientation). Incorporating stratified frequencies into BRAVO could also aid comparison with other variant resources presenting similarly stratified frequencies (e.g., gnomAD, NCBI ALFA). These comparisons would be made more robust if similar approaches to defining/creating strata were used across resources (see Appendix: Precedent from prior and existing genomics resources).

3. Clinical researchers and clinicians to assess variant pathogenicity

A potential use case for the BRAVO variant server is to inform clinical variant interpretation, e.g., to aid in classifying variants as pathogenic or benign. The American College of Medical Genetics and Genomics (ACMG) current guidelines for clinical variant interpretation rely on allele frequency data from "population databases" as an important source of evidence (Richards et al. 2015). For example, the guidelines state, "If a variant is absent from (or below the expected carrier frequency if recessive) a large general population or a control cohort (>1000 individuals) and the population is race-matched to the patient harboring the identified variant, then this observation can be considered a moderate piece of evidence for pathogenicity" (Richards et al. 2015, p 414). Notably, allele frequency *greater* than expected given the disorder, or over 5%, can support a benign interpretation—a point on which the guidelines do not require a matched subpopulation. The current lack of diversity in genetic databases to date means there is a lack of frequency information in non-European ancestry populations and therefore increased difficulty determining pathogenicity of variants detected in patients and families of non-European ancestry. Stratified frequencies on BRAVO could therefore fill an important knowledge gap in ancestry-specific frequencies in a clinical context. Specifically, a variant may be very rare when frequency is computed over the entire TOPMed population, but common when focusing on a specific strata. Therefore using TOPMed-wide versus stratum-specific allele frequencies may lead to different variant classifications.

Similarly, bioinformatics researchers may use databases of allele frequencies at scale to generate predictions of variant pathogenicity (e.g., de Andrade et al. 2018, Jagadeesh et al. 2019). Such measures of predicted pathogenicity are subsequently used in other genomic research and genetic epidemiology, e.g., to filter and weight variants in association analyses.

Caveats for these benefits to clinical variant interpretation include:

1. It is still an open research question how useful stratified frequencies are for clinical variant interpretation.
    - The TOPMed DCC is aware of eight direct inquiries to date from clinician-researchers or clinical laboratories seeking more information on rare variant carriers in BRAVO. Notably, these inquiries typically request phenotype information related to the disease of interest rather than race/ethnicity or genetic ancestry.
2. TOPMed is a heterogeneous research collection, including population-based cohorts, case-control, and family-based studies. Variant interpretation guidelines (e.g., Richards et al. 2015) require distinguishing between population- and disease-based databases, which may complicate the use of TOPMed as a source of evidence.

Concerns

Despite the above benefits, there are also potential concerns with public-facing stratified frequencies, which we group into four main categories below:

- Re-identification

Re-identification in this context means determining whether or not an individual participated in a given research study, based on genetic data of the individual. This information could be stigmatizing or otherwise compromise participant privacy, especially if study participation reveals sensitive phenotypic information about the individual. TOPMed studies from defined geographic areas and/or comprising families may have heightened concerns about re-identification.

Allele frequencies are a form of Genomic Summary Result (GSR), for which NIH data sharing policies have evolved over the past two decades. Initially GSR were made publicly available in NIH-designated repositories. GSR were later moved out of public access following a demonstration by Homer et al. (2008) of the ability to determine whether or not an individual contributed to a study based on access to allele frequencies *and* access to individual-level genetic data from the individual. After much deliberation, the NIH GSR sharing policy was revised again in 2018 such that unless a study designated as "sensitive," GSR could be made publicly available—i.e., unrestricted access (see NOT-OD-19-023). Other publications studied the risk of re-identification using GSR but assuming that results from a genetic association study are available (i.e., including effect size estimates). For example, Rice and Lumley (2009) showed that a phenotype prediction can be constructed for an individual based on their genetic data and effect size estimates from GWAS. However, relevant to our question is the approach using allele frequencies rather than effect estimates. Notably, in either case, re-identification risks are expected to decrease with the size and genetic heterogeneity of the research cohort (Homer et al. 2008; Visscher and Hill 2009; Bacanu 2017), suggesting mitigated risk if the sample size of the strata is large enough.

- Reputational harm

Aside from the privacy risk of re-identification, there are potential reputational or dignitary harms in providing stratified frequencies. Dignitary harms can be defined as those that "undermine the perceived value and worth of the group in the eyes of others and the group itself," (McGregor 2010) and are distinct from individual-level risks of re-identification or other privacy breach. For instance, pathogenic variants that are present only in certain subgroups/strata may stigmatize populations. There is also a potential reputational harm to study investigators and institutions if study participants learn that pathogenic variants may be present in their population but the information is not being returned to affected individuals. The likelihood and potential magnitude of these risks is not well known, however.

- Reification of race

Stratification may encourage or reinforce genetic or biological notions of race. Specifically, presenting genetic information (e.g., variant frequencies) using racial or ethnic groupings suggests those groupings are defined or distinguished by genetic differences, whereas there is overwhelming evidence to the contrary (ASHG 2018). This practice would also not meet the call to promote anti-racism in science (Yudell et al. 2020). Stratification based on genetic ancestry rather than race or ethnicity may still run the same risk. For example, prior studies have shown that genetic ancestry, when conceptualized at the continental level, can still be mapped onto common notions of race (Fujimura and Rajagopalan 2011). Statistical methods that do not

require creation of discrete groups based on either genetic or social definitions may mitigate this risk; however, further empirical ELSI research is needed to evaluate this claim.

- Consent issues

Studies should evaluate whether presenting stratified frequencies is consistent with participant consent. For example, some consents may not allow for population genetics research, or may only allow for research related to a specific disease. This raises the issue of whether presenting stratified frequencies on BRAVO would constitute population genetics research and whether these frequencies can be used for population genetics or other research beyond the scope of participant consent. Active (versus legacy) TOPMed studies also have evolving and potentially multiple iterations of consent to consider, especially for studies spanning multiple recruitment sites and institutions.

The TOPMed ELSI Committee has previously applied a "publication analogy" to determine whether downstream data uses (e.g., of summary data) may be constrained by individual-level consent—including contribution to the BRAVO server—see prior summary documents on the TOPMed BRAVO variant server and the TOPMed imputation reference panel. Briefly, the publication analogy is the concept that once data are published in a journal article, they enter the public domain and may be used beyond the limits defined during the original participant consent process. The publication analogy could reasonably be applied to presentation of stratified allele frequencies in BRAVO, as it was previously to TOPMed-wide frequencies in BRAVO. Notably, according to current NIH GSR sharing policies, full GSR from "sensitive" studies remain under controlled access and do inherit the individual-level consents, which seemingly contradicts the publication analogy. Still, some limited aspects of GSR from sensitive studies may yet enter the public domain separately from individual-level consent—e.g., a table of top "hits" (most strongly associated variants with phenotype) in a publication, or top hits with redacted direction of effect in the dbGaP Genome Browser, suggesting that sensitivity designation is concerned with the potential risks of re-identifiability and of reputational harm, rather than with informed consent (see NOT-OD-19-023).

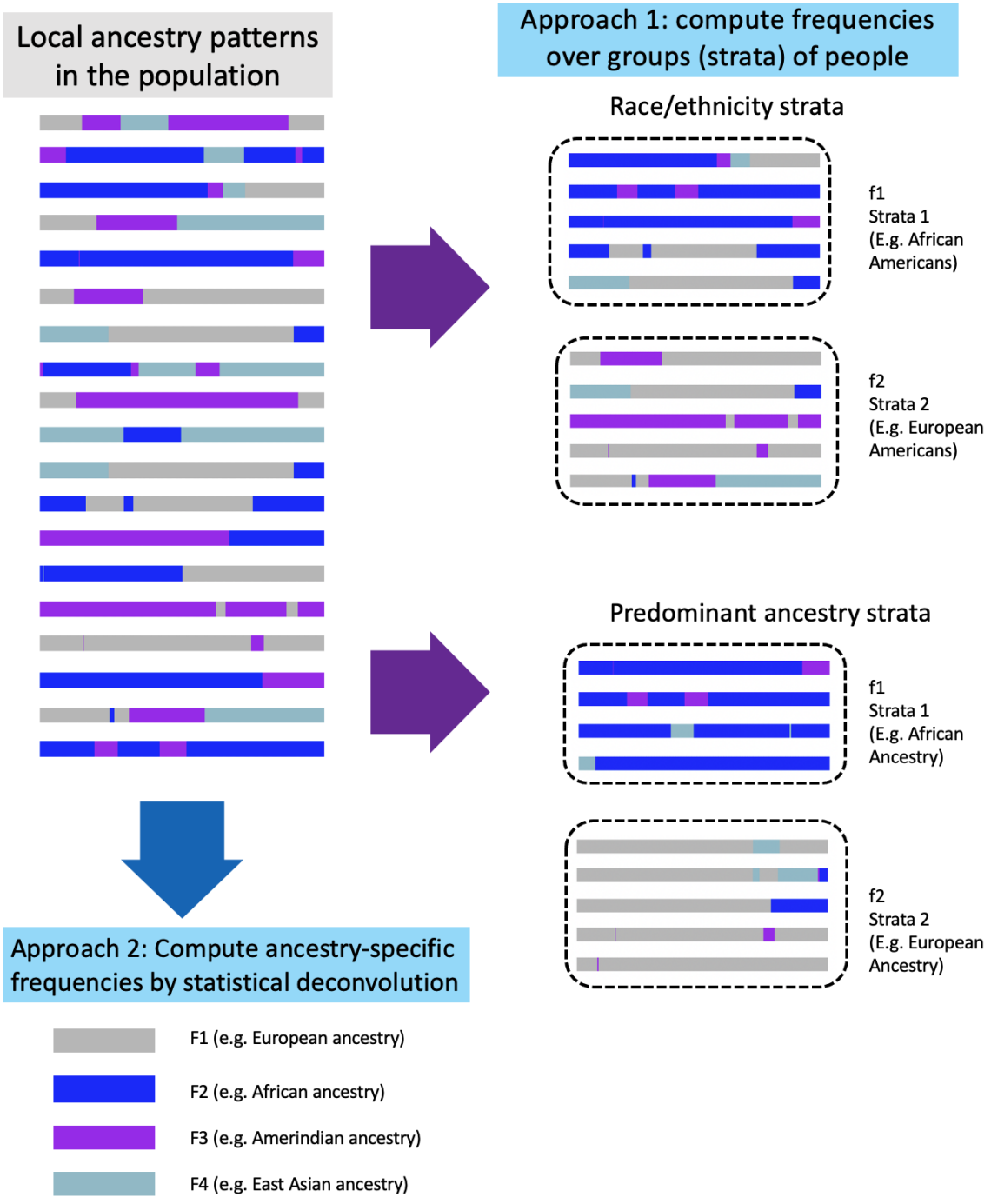
Ultimately, participant informed consent processes for large-scale genomic research are complicated (McGuire and Beskow 2010), including by potential downstream data uses often unconceived or unknowable at the time of participant recruitment. While it is unlikely investigators solicited participant preferences about summary results sharing at the time of recruitment, that does not absolve investigators and institutions from considering whether such downstream uses are consistent, or at least not inconsistent, with participant wishes and understandings at the time of consent. Therefore, studies should consider the consent-related implications noted above.

Considerations for implementation

There are multiple approaches to creating or defining strata, each of which has advantages and disadvantages. Furthermore, the specifics of proposed strata will likely influence whether TOPMed study investigators agree to have their study data included. Here we describe some potential routes for defining strata, illustrated in Figure 1 and summarized in Table 1, and note considerations for different options. In the Conclusion and Recommendations section, we

ultimately recommend ancestry-specific allele frequencies by statistical deconvolution (Approach 2 in Figure 1) as the preferred approach in TOPMed given the advantages and disadvantages.

Figure 1: Visualization of approaches for computing stratified allele frequencies.



In brief, allele frequencies can be computed based on an approach in which *individuals are grouped* according to common characteristics, e.g. by social definitions of race/ethnicity or

based on genetic ancestry patterns (Approach 1 in Figure 1). Alternatively, allele frequencies can be computed without defining groups, but rather by first inferring the genetic ancestral background of each individual in the data and then using this inference to *deconvolve* frequencies (Approach 2 in Figure 1). The latter approach acknowledges that every group of individuals is a mixture of genetic ancestries, and once the distribution of these ancestries within this mixture is known, one can infer frequencies of variants in each of these ancestries.

We propose to estimate ancestry-specific allele frequencies using a statistical deconvolution method and based on local genetic ancestry inference. In more detail, TOPMed investigators at the IRC previously performed local and global genetic ancestry inferences for TOPMed genomes. They condensed the 53 Human Genome Diversity Project (HGDP) reference populations into seven "super populations" to assign genetic ancestries to TOPMed samples. Specifically, each genome is divided into segments (local ancestry intervals, each encompassing a range of haplotypes) that are categorized to one of these ancestries (local ancestry at the interval). Global ancestries are averages of these local ancestries across each genome (see freeze 8 local ancestry README for details). It is possible to compute ancestry-specific allele frequencies using statistical algorithms applied across all available TOPMed participants by incorporating local ancestry information. Advantages and disadvantages of using this approach are:

Advantages

- Using genetically-inferred measures avoids harmonizing demographic categories across TOPMed
- Avoids grouping individuals into categories, which are necessarily imprecise
- Mitigates reidentification concerns because the mixture distribution of ancestral allele frequencies varies across the genome (i.e. by local ancestry)
- Maximizes inclusion of individuals—i.e. does not require meeting a threshold of global ancestry proportion, or having non-missing race/ethnicity information that fits into a harmonization (e.g., US-based) framework.
- Due to the low risk of re-identification, this approach does not require specifying a minimum number of TOPMed studies or of individuals in order to report a specific strata, unless a specific continental ancestry is uniquely and completely represented by a specific TOPMed study.
    - See the effective sample size figure in the Appendix for an overall sense of sample sizes across the seven HGDP-defined super populations.

Disadvantages and potential mitigations

- The availability and selection of reference populations will constrain the possible ancestries. Currently, we propose to use the 7 continental ancestries represented in the HGDP reference. While more granular ancestral definitions may be useful (e.g., sub populations of South Asians), appropriate reference data is currently lacking. In the future, we may re-evaluate the ancestries used based on data availability. We also note that all stratified allele frequency approaches are affected by the availability of reference data.

- Allele frequency estimates using statistical deconvolution may be less precise, compared to standard estimates, in some settings (e.g., rare variants, or even moderately common variants on haplotypes that are only rarely available from a specific ancestry). Therefore, we will not be able to confidently share stratified variant frequencies for rare variants.
    - **Mitigation:** Continue to provide global TOPMed frequencies for rare variants.
- This approach would not accommodate a study interested in sharing study-specific allele frequencies, e.g. for studies from a unique founder population.
    - **Mitigation:** We recommend an additional mechanism that would allow such sharing, e.g. see recommendation #2 in Conclusion and Recommendations.

Table 1 summarizing three types of stratified frequencies, their advantages and disadvantages. For further details on the two grouping approaches, see Appendix: Advantages and disadvantages to grouping approaches.

	Ancestry-specific allele frequencies by statistical deconvolution	**Ancestry-specific allele frequencies by grouping individuals according to ancestral patterns**	**Race/Ethnicity group specific allele frequencies**
Grouping approach	No	Yes	Yes
All participants can contribute to analysis?	Yes	Only those who are "mostly" from one ancestry (i.e. predominant ancestry)	Only those with reported race/ethnicity
Risk of re-identification?	Very low; re-identification methods do not currently exist	Risk exists for groups with small sample size	Risk exists for groups with small sample size
Reification of race as a biological variable?	Low	Medium (can be conflated with genetic ancestry)	High

Conclusion and Recommendations

In summary, **stratified allele frequencies are useful for quality control, design of replication studies, population genetics research, and clinical variant interpretation**. The latter use of stratified variant frequencies is still under development, in that the effect of genetic ancestry on pathogenicity is still not studied (e.g., gene x gene interaction), and it is not clear whether ancestry-specific frequencies are indeed more useful compared to frequencies estimated on groups defined using social constructs such as race or ethnicity.

The risks associated with sharing allele frequencies outside the context of a specific GWAS (and therefore a specific phenotype) are small and become even smaller when combining together populations from multiple contributing studies. In addition, using statistical

methods that do not require grouping means it is not necessary to define and adhere to a minimum number of participants or studies per group/strata.

We recommend:

1. **Providing ancestry-specific allele frequencies by statistical deconvolution, in addition to TOPMed-wide frequencies**. This approach has the benefits of avoiding reification of race, maximizing use of available data, and mitigating risks of re-identification and reputational harms (see Table 1). An additional benefit is the potential to represent populations that are not generally available in a predominant ancestry framework (e.g., Amerindian ancestry). We recognize the clinical utility of this novel approach is not yet known, e.g. for clinical variant interpretation. However, we contend this approach is preferable to the imprecision of trying to "race-match" patients to reference databases, when both patient and reference genomes are a mosaic of local ancestry patterns.

2. **Individual studies and cohorts may choose to additionally participate as a unique strata.** Studies may proactively engage with the IRC if they are interested to contribute to a unique strata, in addition to participating in ancestry-specific allele frequencies as described above. For example, a founder or other distinctive population (e.g., Amish) may elect to provide population-specific frequencies from their study. As with the ancestry-specific frequencies, studies would consider on a case-by-case basis whether this would be appropriate and desirable for their study.

3. **Study PIs and representatives should carefully consider participant consent when deciding whether to contribute to stratified allele frequencies in BRAVO.** Specifically, a risk of sharing any stratified frequencies is that they lend themselves to population genetics research, which may be outside the scope of participants' consents. Furthermore, some study populations may have concerns or sensitivities about such research even if not explicit in informed consent language.

4. **TOPMed stratified frequencies should be transparently and consistently described and used across resources and publications.** The approach used to create ancestry-specific frequencies should be clearly described on the BRAVO website. Other variant resources created using TOPMed data (e.g., annotation servers) should follow a similar approach wherever feasible.

5. **The TOPMed program should revisit and revise the definition and composition of strata in TOPMed as needed moving forward.** For example, the incorporation of additional TOPMed studies may suggest new approaches or strata, or revised definitions of existing strata, i.e., as new datasets are included in TOPMed WGS freezes. Proposals for future revision to TOPMed strata should first be addressed to the TOPMed Executive Committee.

Ultimately, whether and how stratified frequencies are added to BRAVO is a combination of study-specific ("bottom up") and TOPMed program ("top down") decisions and deliberations. Here we have presented "points to consider" for decision-makers at both the study and program levels, articulating potential benefits and risks of stratification in general and of alternate approaches to defining strata. In closing, the Committee concludes that potential benefits of providing stratified frequencies likely outweigh potential risks for most TOPMed studies, though each study will ultimately need to make their own determination.

References

ASHG Denounces Attempts to Link Genetics and Racial Supremacy (2018). In American Journal of Human Genetics (Vol. 103, Issue 5, p. 636). Cell Press. https://doi.org/10.1016/j.ajhg.2018.10.011

Bacanu S-A (2017). Sharing extended summary data from contemporary genetics studies is unlikely to threaten subject privacy. PLoS ONE 12(6): e0179504

De Andrade KC et al. (2018). Variable population prevalence estimates of germline TP53 variants: A gnomAD-based analysis. Human Mutation 40(1), 97-105

Fang et al. (2019). “Harmonizing Genetic Ancestry and Self-Identified Race/Ethnicity in Genome-Wide Association Studies.” American Journal of Human Genetics, 105(4), 763-772

Fujimura JH, Rajagopalan R (2011). Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research. Social Studies of Science 41(1).

Homer N et al. (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genetics 4(8)

Hunter-Zinck H et al. (2020). Measuring genetic variation in the multi-ethnic Million Veteran Program (MVP). bioRxiv (preprint). Available at https://www.biorxiv.org/content/10.1101/2020.01.06.896613v1.full.

Jagadeesh KA et al. (2019). S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. Nature Genetic 51(4), 755-763

Lumley T, Rice K (2010). Potential for Revealing Individual-Level Information in Genome-wide Association Studies. JAMA. 303(7):659–660. doi:10.1001/jama.2010.120

McGregor J (2010). Racial, ethnic, and tribal classifications in biomedical research with biological and group harm. American Journal of Bioethics, 10(9), 23–24

McGuire AL, Beskow LM (2010). Informed consent in genomics and genetic research. Annu. Rev. Genomics Hum. Genet. 11, 361–81

Nelson SC et al. (2018). A content analysis of the views of genetics professionals on race, ancestry, and genetics. AJOB Empir. Bioeth. 9, 222–234

Popejoy AB et al. (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am. J. Hum. Genet. 107, 72–82

Richards S, Aziz N, Bale S et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405–423

Taliun D, Harris DN, Kessler MD et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299

Visscher PM, Hill WG (2009). The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis. PLoS Genet 5(10): e1000628

Yudell M, Roberts D, DeSalle R, & Tishkoff S (2020). NIH must confront the use of race in science. Science, 369(6509), 1313–1314. https://doi.org/10.1126/science.abd4842

Appendices

Advantages and disadvantages to grouping approaches

Self-reported racial and/or ethnic categories

Advantages

- Racial or ethnic categories may map to patient or research participant categories for which BRAVO users are seeking information.

Disadvantages

- There may be sporadic missingness leading to the exclusion of participants, though values could perhaps be imputed, e.g., via HARE (Fang et al. 2019), leading to a combination of self-report with genetic-based information.
- Potential reification of these categories as genetic or biological (see Concerns: Reification of race).
- Race/ethnic categories are subjective and defined by individuals and institutions according to social and historical norms. Anecdotally, a specific individual self-identifying (or lab/clinician-identified) with one race/ethnic group may be genetically more similar to TOPMed individuals from a different race/ethnic group category on BRAVO. Indeed these classifications lack definitional clarity and consistency in research and clinical contexts (e.g., Popejoy et al. 2020; Nelson, Yu, et al. 2019).
- Harmonized race and ethnicity are not available across all of TOPMed and harmonization is complicated by the presence of non-US based studies that are unlikely to use administrative race/ethnicity categories common in the US.

Predominant ancestry groups

Based on admixture analysis, one can estimate the ancestral make-up of TOPMed individuals and assign some of the individuals into “predominant ancestry” groups. For example, an individual with 90% of higher European ancestry will be classified to the European ancestry group. This approach is used in TOP-LD (see Appendix: TOP-LD) and the Million Veteran Program (Hunter-Zinck et al. 2020).

Advantages

- Using genetically-inferred measures avoids harmonizing demographic categories across TOPMed.

Disadvantages

- The availability and selection of reference populations will constrain the TOPMed groupings that are possible.
- Individuals will be omitted if they fail to meet the (arbitrary) threshold for a group, which could lead to underrepresentation of specific (e.g., admixed) populations.
- This approach shares the same earlier concern of reifying race, especially given that predominant ancestry groups could be perceived as discrete and homogeneous

TOPMed effective sample size

The bar plot below shows effective sample size of the seven HGDP “super populations” across TOPMed freeze 8, with the caveat that the effective sample size may decrease if cohorts opt out. Notably, these sample sizes do not reflect discrete, fixed groups of samples in our recommended approach of estimating ancestry-specific allele frequencies using a statistical deconvolution method and based on local genetic ancestry inference. Instead, they provide the summation across all fractions of genomes from each of the ancestries. The power for estimating allele frequencies in any specific ancestry increases with higher effective sample size.

Figure S1. Effective sample size

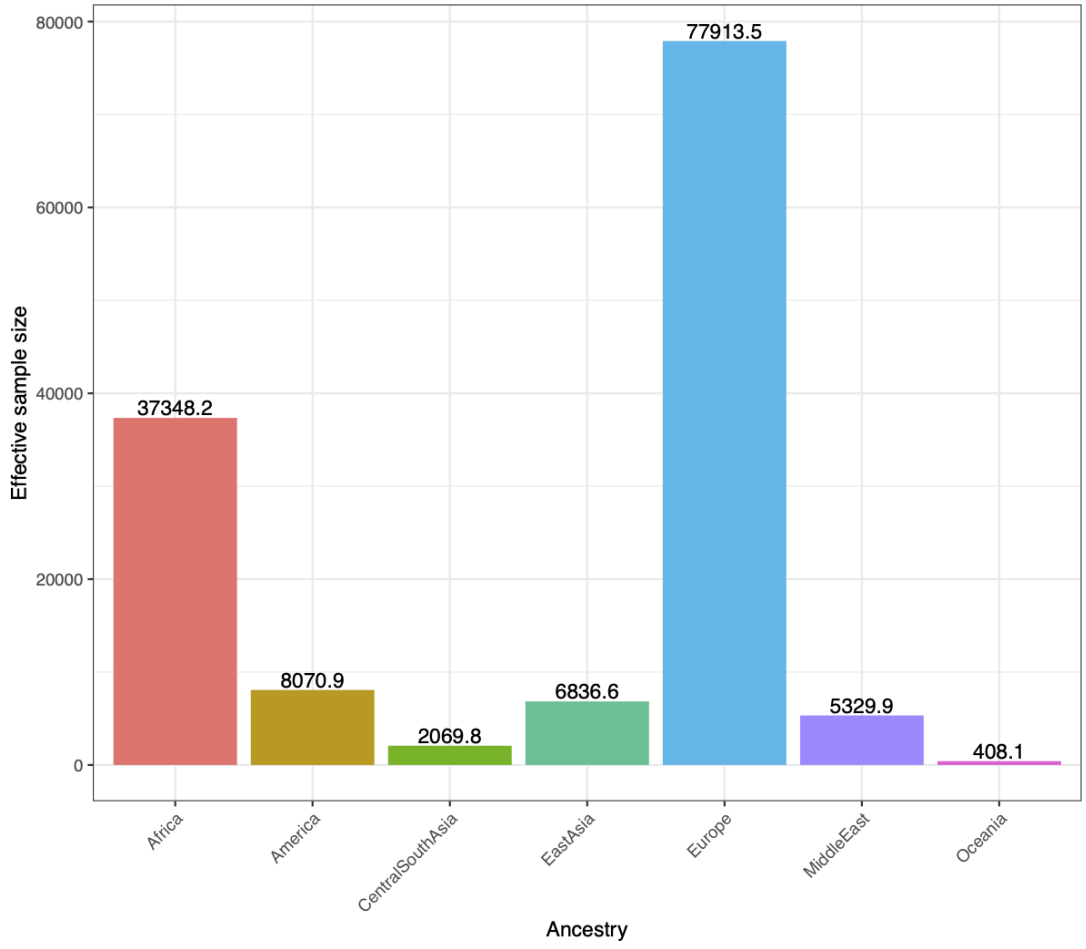
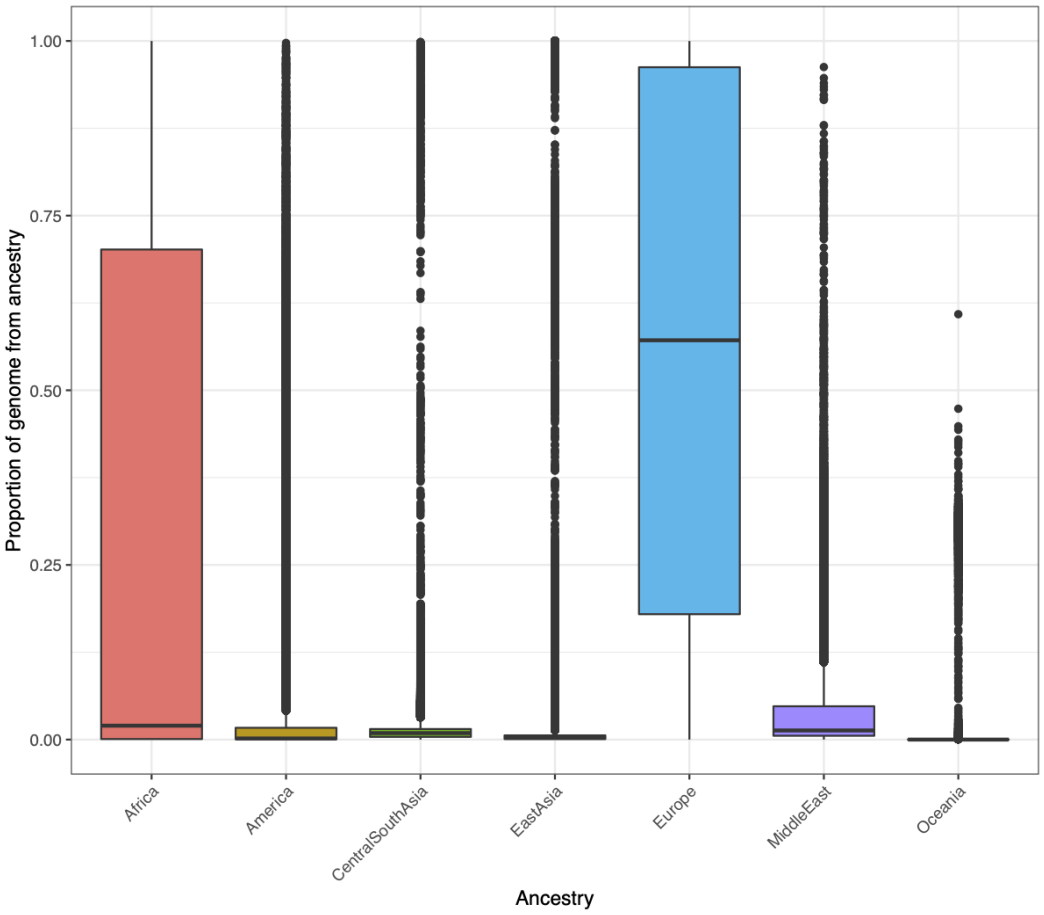


Figure S2. Ancestry proportions

The boxplots below show genome-wide ancestry proportions across TOPMed freeze 8, again using the HGDP super populations as reference. Most individuals have only a small fraction of genetic ancestries associated with super populations from America, Central and South Asia,

East Asia, Middle East, and Oceania. Statistical deconvolution algorithms allow for estimating frequencies from these ancestries while methods that group individuals (either by reported race/ethnic group or by pedominant ancestry) would not.



Precedent from prior and existing genomics resources

1. [1000 Genomes Project]

   5 “super populations”:
   - AFR, African
   - AMR, Ad Mixed American
   - EAS, East Asian
   - EUR, European
   - SAS, South Asian

2. [gnomAD]

Population	Description	Genomes
afr	African/African-American	20,744

ami	Amish	456
amr	Latino/Admixed American	7,647
asj	Ashkenazi Jewish	1,736
eas	East Asian	2,604
fin	Finnish	5,316
nfe	Non-Finnish European	34,029
mid	Middle Eastern	158
sas	South Asian	2,419
oth	Other (population not assigned)	1,047

3. ALFA (Allele Frequency Aggregator)

- The ALFA resource presents variant information aggregated across dbGaP studies that aren't designated as "sensitive" with respect to GSR sharing.

Name	Population Code	Description
African Others	AFO	Individuals with African ancestry
African American	AFA	African American
African	AFR	All Africans, AFO and AFA Individuals
European	EUR	European
Latin American 1	LAC	Latin American individuals with Afro-Caribbean ancestry
Latin American 2	LEN	Latin American individuals with mostly European and Native American Ancestry
South Asian	SAS	South Asian
East Asian	EAS	East Asian (95%)
Asian	ASN	All Asian individuals (EAS and OAS) excluding South Asian (SAS)
Other Asian	OAS	Asian individuals excluding South or East Asian
Other	OTR	The self-reported population is inconsistent with the GRAF-assigned population

4. TOP-LD

A public-facing LD server based on TOPMed data under development by Paul Auer and Yun Li - see also TOPMed paper proposal 10711. Ancestry groups are defined in four TOPMed studies (BioMe, MESA, JHS, and WHI) using RFMix results with 1000 Genomes Project and HGDP as references, only considering samples with >90% ancestry from a single population, and removing related individuals:

- African ancestry
- East Asian ancestry
- European ancestry
- South Asian ancestry

Document history

- May 24, 2021 - approved by TOPMed Executive Committee for distribution to TOPMed study PIs
- April 12, 2021 - approved by TOPMed ELSI Committee
- April 2, 2021 - revision circulated to TOPMed ELSI Committee for review
- Feb 11, 2021 - circulated to TOPMed ELSI Committee for review
- Jan-Feb, 2021 - drafted by ELSI Committee co-conveners, Sarah Nelson and Tamar Sofer
- Aug-Dec, 2020 - initial discussions in the ELSI Committee with IRC representative Albert Smith